A Digital Coach That Provides Affective and Social Learning Support to Low-Literate Learners

Dylan G.M. Schouten, Fleur Venneker, Tibor Bosse, Mark A. Neerincx, and Anita H.M. Cremers

Abstract—In this study, we investigate if a digital coach for low-literate learners that provides cognitive learning support based on scaffolding can be improved by adding affective learning support based on motivational interviewing, and social learning support based on small talk. Several knowledge gaps are identified: motivational interviewing and small talk must be translated to control rules for this coach, a formal model of participant emotional states is needed to allow the coach to parse the learner's emotional state, and various sensors must be used to let the coach detect and act on this state. We use the situated Cognitive Engineering (sCE) method to update an existing foundation of knowledge with emotional models, motivational interviewing, and small talk theory, technology, and a new exercise in the volunteer work domain. We use this foundation to create a design specification for an Embodied Conversational Agent (ECA) coach that provides cognitive, affective, and social learning support for this exercise. A prototype is created, and compared to a prototype that only provides cognitive support in a within- and between-subjects experiment. Results show that both prototypes work as expected: learners interact with the coach and complete all exercises. Almost no significant differences are found between the two prototypes, indicating that the affective and social support were not effective as designed. Potential improvements are provided for future work. Results also show significant differences between two subgroups of low-literate participants, and between men and women, reinforcing the importance of using individualized support measures with this demographic.

Index Terms—Affective computing, computer aided instruction, electronic learning, emotion recognition

1 Introduction

In earlier studies, we have highlighted the problems that people of low-literacy encounter when trying to participate in information societies [1], [2]. Low information (reading and writing) and communication (speaking and understanding) skills cause participation issues that can be of a cognitive nature (skill application and general societal knowledge), affective nature (emotional responses like shame and fear, and low self-efficacy), and/or social nature (motivation to participate and trusting peers and teachers). We want to address these issues by designing interactive, situated societal participation learning that is grounded in crucial practical situations, which are real-life scenarios that describe the skills and knowledge needed for independent societal participation [3]. The aim is to make learning more effective, which means making the learning process more accessible (by removing or lowering barriers to entry) and making the learning experience more positive (ensuring that learners both can and want to interact with the learning), thereby supporting learners in reaching desired learning outcomes [1]. Specifically, we are designing the system *VESSEL*: a *Virtual Environment to Support the Societal participation Education of Low-literates*. VESSEL is envisioned as a set of interactive exercises grounded in the aforementioned crucial practical situations, and an autonomous, rules-driven *Embodied Conversational Agent* (*ECA*) coach that helps low-literate learners carry out these exercises by offering cognitive, affective, and social learning support. Fig. 1 shows a schematic VESSEL design.

We use the *situated Cognitive Engineering* method in the VESSEL development process (*sCE*, see [5], [6]). This iterative software design and development method consists of three stages. In the *foundation* stage, relevant *operational demands* (actors, activities, and context-of-use), *human factors data* (theory relevant to user-system interaction), and *technology* are collected. In the *specification* stage a *requirements baseline* is created, consisting of *functional requirements* (the system's intended functionality), *claims* (hypotheses that describe the system's intended effects), *system objectives* (the system's operational or domain goals), and *use cases* (action sequences that describe the system's ideal working procedure). In the *evaluation* stage, this requirements baseline is experimentally validated.

In prior studies we have designed, developed, and evaluated two VESSEL prototypes. The first prototype [4] was a proof-of-concept consisting of four information-and-communication-skill exercises (easy and hard variants of 'online banking' and 'service desk conversation' exercises)

- *D.G.M. Schouten is with the Interactive Intelligence Group (II), Faculty Electrical Engineering, Mathematics, and Computer Science, TU Delft. Mekelweg 4, Delft 2628 CD, Netherlands. E-mail: dylan.schouten@gmail.com.*
- *F. Venneker is with the Faculty of Science, University of Amsterdam, Science Park 904, Amsterdam 1098 XH, Netherlands. E-mail: fleurvenneker@msn.com.*
- *T. Bosse is with the Behavioural Informatics Group, Faculty of Sciences, Vrije Universiteit Amsterdam, De Boelelaan 1081a, Amsterdam 1081 HV, Netherlands. E-mail: t.bosse@vu.nl.*
- *M.A. Neerincx is with TNO and the II, Faculty Electrical Engineering, Mathematics, and Computer Science, TU Delft. Mekelweg 4, Delft 2628 CD, Netherlands. E-mail: mark.neerincx@tno.nl.*
- *A.H.M. Cremers is with TNO, Kampweg 5, Soesterberg 3769 DE, Netherlands. E-mail: anita.cremers@tno.nl.*

Manuscript received 15 Dec. 2016; revised 2 Apr. 2017; accepted 12 Apr. 2017. Date of publication 2 May 2017; date of current version 28 Mar. 2018. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below. Digital Object Identifier no. 10.1109/TLT.2017.2698471

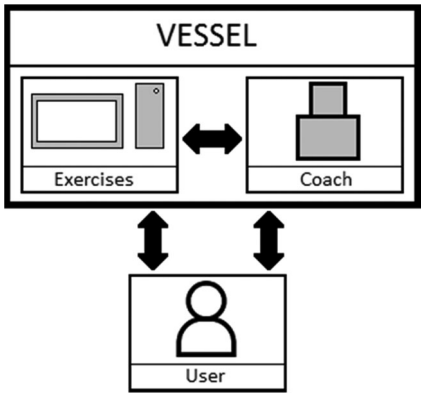


Fig. 1. VESSEL design. System interactions are indicated with arrows: the user performs exercises, the coach monitors exercise state and user-system interaction, and the coach supports the user [4].

and an ECA coach offering three kinds of learning support: cognitive support based on scaffolding (a learning support method that provides the right amount of help at the right time, [7]), affective support based on motivational interviewing (a counseling technique focused on enacting behavioural change, [8]), and social support based on small talk (a form of social interaction that is important to building interpersonal trust, [9]). All support was given as pre-recorded spoken utterances, controlled by a Wizard-of-Oz operator (see [10]). We evaluated this prototype to test the general applicability of cognitive, affective, and social support offered by an ECA coach. Results showed that the ECA coach improved the learning experience in all facets (cognitively, affectively, and socially), and raised learner self-efficacy regarding challenging online banking situations. Based on these positive results, a design specification for coach-driven cognitive learning support was drafted, and translated into a second prototype [11]. This work consisted of three challenging online banking exercises, and an ECA coach offering cognitive learning support based on formal scaffolding theory (see [12]) while following strict speech recognition rules. We evaluated this prototype (still controlled via the Wizard-of-Oz method) to test our claims of benefit with regard to cognitive support. Results showed that learner self-efficacy regarding challenging online banking was again raised, and that the formalized coach did not negatively impact the learning experience: expectations were that participants would try to interact with the coach as if it was human (i.e., asking complex questions and expecting the coach to have an answer for every situation), and that the coach's limited knowledge and strict speech recognition could cause difficulty and frustration. But this did not happen.

Now, we want to extend the VESSEL specification by also incorporating affective and social support into the design specification, thus bringing system functionality in line with the envisioned functionality from the proof-of-concept. However, trying to do so illustrates two important knowledge questions about VESSEL's ECA coach. First, we need to know how to design affective and social learning support for low-literate learners, in particular how to translate motivational interviewing theory (for affective support) and small talk theory (for social support) into support rules for the coach. Second, affective support specifically depends on understanding the learner's emotional state. We need to know what technology would allow the ECA coach to perceive and react to learner emotions, and which emotional models we can use to categorize these. As in [11], we can answer these questions by incorporating new theory into the sCE foundation of VESSEL. We update operational demands by designing one or more new scenario-based exercises that demand cognitive, affective, and social support. We update human factors knowledge by incorporating theory on motivational interviewing, small talk, and emotional models. We update technology by describing both current technology for autonomous emotion detection, and the envisioned role of this technology in VESSEL.

In summary, in this work we aim to design and evaluate a third VESSEL prototype that offers cognitive, affective, and social learning support. Four steps are needed. First, we expand the VESSEL foundation as described. Second, we refine the VESSEL design specification by operationalizing the foundation theory into comprehensive coach behavioural rules, updating the requirements baseline, and writing new uses cases. Third, we design and develop our third VESSEL prototype on the basis of this specification. Finally, we experimentally evaluate the prototype with low-literate learners. Specifically, we investigate how the new prototype affects the cognitive, affective, and social learning experience and learning outcomes of a volunteer work learning exercise, compared to our previous prototype (see [11]). This lets us answer the following research questions:

Q1. *Design*: How can we create a design specification for VESSEL that incorporates rules for cognitive, affective, and social learning support provided by an ECA coach?
Q1a. Which emotional models, motivational interviewing rules, small talk scenarios, and measurement methods are needed to create these rules?
Q1b. Which functionalities, interaction methods, and appearances should the ECA coach have to reflect this?
Q2. *Evaluation*: Does an ECA coach created in accordance with this specification result in a higher learning effectiveness for low-literate learners than an ECA coach that incorporates only formalized cognitive learning support?

The structure of this paper is as follows. In Section 2, the sCE foundation is updated to address the knowledge gaps: what coach behaviour rules can be derived from motivational interviewing and small talk theory, which formal models of emotion can be used by the ECA coach, and what technological options are there for autonomous emotion detection? This information is incorporated into the sCE foundation in Section 3. In Section 4, the new VESSEL prototype is designed and developed. Section 5 describes the design and setup of the experiment created to evaluate the effectiveness of the prototype, and Section 6 presents evaluation results. Finally, Section 7 presents conclusions and directions for future work.

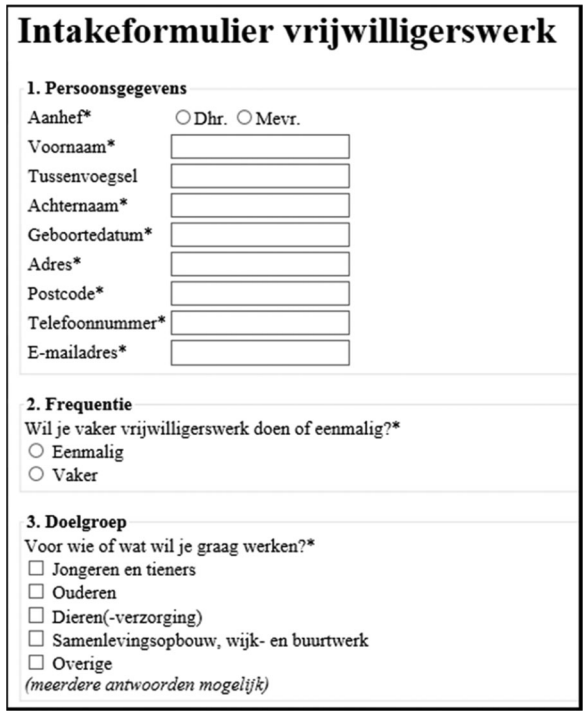
2 Foundation

2.1 Operational Demands: Exercises

To provide the right context-of-use for the envisioned cognitive, affective, and social support coach, exercises are needed that pose cognitive, affective, and social challenges

Fig. 2. The two appearance options for the ECA recruiter.

(in tandem). None of our previous exercises (see [4], [11]) meet this demand. We have chosen to design a new exercise, based on the crucial practical situation 'registering for volunteer work'. The exercise consists of two parts. In the 'form' part, learners must fill out an 'intake form' for volunteer work. The form has a section for demographic information, and four sections that categorize the learner's wishes with regard to volunteer work: frequency, target demographic(s), target area(s), and useful skills possessed by the learner. This part of the exercise tests reading and language comprehension, as well as ICT skills, and presents mostly potential cognitive problems (related to vocabulary and comprehension), but also affective ones (willingness to admit interests, uncertainty about what this information is used for). In the 'recruiter' part, learners must speak to an ECA playing the role of a volunteer work recruiter. The recruiter asks a number of questions, drawn from a large set, that reference their choices on the form. Learners talk to the recruiter directly. This part of the exercise tests speaking skills and comprehension of spoken language, and presents mostly potential affective problems (fear and shame about discussing personal desires and limitations) and social ones (speaking to a formal-looking stranger about unfamiliar topics). Combined, we think that cognitive, affective, and social challenges will be presented throughout the exercise, providing room for the coach ECA to support learners in all three areas. Two versions of the exercise have been made: the order of information elements and some of the contents are different in the forms, and the recruiter ECAs are visually slightly different. Fig. 2 shows the two appearances of the recruiter ECA. Fig. 3 shows an excerpt of one exercise form.



Intakeformulier vrijwilligerswerk

1. Persoonsgegevens

Aanhef* ○ Dhr. ○ Mevr.
Voornaam*
Tussenvoegsel
Achternaam*
Geboortedatum*
Adres*
Postcode*
Telefoonnummer*
E-mailadres*

2. Frequentie

Wil je vaker vrijwilligerswerk doen of eenmalig?*
○ Eenmalig
○ Vaker

3. Doelgroep

Voor wie of wat wil je graag werken?*
☐ Jongeren en tieners
☐ Ouderen
☐ Dieren(-verzorging)
☐ Samenlevingsopbouw, wijk- en buurtwerk
☐ Overige
(meerdere antwoorden mogelijk)

Fig. 3. Excerpt of one variant of the intake form, containing questions about: demographic information, frequency of volunteer work, and intended target group to do volunteer work for.

2.2 Human Factors Knowledge

2.2.1 Emotion Models

To design an ECA coach that can give accurate affective support, we need a way to categorize and assess the intensity of learner emotions. Three general approaches to emotional modeling exist: the basic emotions approach, the cognitive appraisal approach, and the dimensional approach [13]. The *basic emotions* approach claims that certain core emotions are biologically based and genetically coded [14], and that emotions have evolved to increase odds of survival. Ekman [15] posits that the emotions anger, disgust, fear, enjoyment, sadness, and surprise have universal facial expressions associated with them; this makes these emotions basic emotions. Emotions without universal facial expressions (such as awe, excitement, shame, and relief) are conceptualized as a blend of these six. Similar models by Plutchik [16] and Parrott [17] categorize all emotions as primary, secondary, or tertiary emotions. The basic emotions approach is useful for VESSEL because it provides a discrete, easy-to-interpret classification of emotions. However, this classification does not allow for differentiation in the intensity of emotions. This might make this approach too broad-strokes to allow for individually tailored emotional support.

The *cognitive appraisal* approach describes an emotional state as a reaction to an arousing situation [18]. The experience of 'emotion' is attributed to physiological changes in the body [19]. By categorizing which physiological changes respond to which emotional reactions, it becomes possible to measure emotions objectively: for instance, the Facial Action Coding System (FACS, [20]) is a systematic analysis of the emotions associated with facial expressions. This possibility makes the cognitive appraisal approach potentially useful for VESSEL. However, appraisals of events and the associated emotional reactions are individually and culturally variable: different people interpret body signals differently, and different cultures consider some emotions as undesirable or unacceptable [21]. Consequently, using this approach necessitates a careful study of the intended user demographic.

Finally, the *dimensional* approach posits that emotions are not independent discrete states, but rather that all emotions are related in a systematic manner [22]. For instance, Russel's

TABLE 1
Categorization of Four Basic ECA Coach Emotions

Emotion	Valence	Arousal	Dominance
Anger	Negative	High	High
Fear	Negative	High	Low
Sadness	Negative	Low	*Either*
Happiness	Positive	*Either*	*Either*

[23] circumplex model of affect classifies emotions on the axes of valence (how positive or negative the emotion is felt as) and arousal (how excited or calm the emotion is felt as). This approach is useful for VESSEL because it allows for emotional responses with different intensities: for instance, a person can be a little bit happy because the weather outside is nice, very happy but relatively calm when spending time with family and friends, or incredibly happy and excited for winning the lottery. However, no single 'best' classification model exists. Both Oliveira et al. [24] and Lewis et al. [25] claim that the valence and arousal dimensions are actually correlated, and cannot be treated as independent. Others have posited that a third dimension, dominance (how dominant or submissive the emotion is felt as) is necessary to adequately describe the emotion space [26], [27].

In our VESSEL design, we apply a combination of the basic and dimensional models. We define that the coach can categorize four basic emotions: anger, fear, sadness, and happiness. Based on [1], we think these emotions will play a role in our volunteer work scenario. Low-literate people experience sadness or anger when confronted with challenging information tasks, like the complex wording of the form or the difficult vocabulary of the conversation partner. They experience fear when confronted with decisions they feel they cannot oversee the scope of, like being asked to commit to volunteer work. They experience happiness when completing challenging tasks, particularly related to literacy. Additionally, these four emotions can easily be categorized using the three dimensional terms used by [27]: valence/pleasure, arousal, and dominance (see Table 1). These two models describe all the affective functionality we want in our coach: a simple categorization of emotions that a digital coach can recognize, and a division in measurable quantities that can be used for decision purposes. In theory, the cognitive appraisal model could be used to fine-tune the Table 1 classification of basic emotions to low-literate people, and result in a more accurate description of how these emotions are expressed (how strongly, and in what ways). We leave this time- and labor-intensive adaptation out of our current model, and defer it to future work.

2.2.2 Motivational Interviewing

Motivational interviewing is originally a counseling technique aimed at enhancing an individual's intrinsic motivation to change behaviour [8], [28], [29]. The technique has also been used to provide learning support, by making learners feel good about the process, and reframing and reinforcing positive self-efficacy information [30], [31]. The motivational interviewing process consists of three strategies: affirmation, awareness, and alternatives [29]. *Affirmation* aims to establish empathy between counselor and client. This is often combined with *reflective listening* [8], [32] to put the focus of the conversation on the client's perspective, not the counselor's, motivating the client to explore their own thoughts. *Awareness* aims to help clients become aware of their problem through their own reasoning process. The *alternatives* strategy focuses on helping clients evaluate alternatives to their current situation. Sobell and Sobell [31] add two more strategies: *normalizing* aims to communicate to clients that many other people share their problems and their difficulties to change, and *self-efficacy supporting* focuses on raising the client's self-efficacy about being able to make the change.

In VESSEL, we use four of these strategies to formalize the coach's affective support. The awareness and alternatives strategies are most designed for behavioural change therapy, and are therefore less useful in a learning support setting. The remaining strategies (reflective listening, normalizing, affirmation, and self-efficacy supporting) are used to create a four-tiered model of motivational interviewing utterances. By using the four strategies in the orders presented, the coach provides affective support in a standardized way. We further specify that the coach can identify learner emotional states at three levels of accuracy: *General*, *Specific*, and *Very Specific*. If the coach identifies that the learner is in some negative-valence emotional state, but cannot the exact state, the General level of support is used. If the coach can identify the exact emotional state, the Specific level is used. If the coach can even estimate what the exact trigger is for this emotional state (a particular difficult exercise element or challenge), the Very Specific level is used. Table 2 provides an overview of this model, with example utterances for each category and level.

2.2.3 Small Talk

According to Bickmore and Cassell [33], an essential aspect of human-system interaction is building trust between user and application. They show that, in interactive systems, "*. . .embodied conversational agents are ideally suited for this task* [i.e., building trust] *given the myriad cues available to them for signaling trustworthiness*" (p. 396). In learning, trust makes learners more receptive to teacher suggestions, and motivates learner persistence [33], [34]. Small talk is often used to establish trust. Cassell and Bickmore [9] show that small talk leads to trust-building in three ways. First, small talk establishes *solidarity*, demonstrates reciprocal appreciation, and avoids 'face threat', both because the speakers show interest in one another and because the conversation is kept on a safe level of depth. Second, it establishes *familiarity* and common ground, because speakers discuss a clearly established and accessibly topic. Third, small talk increases *coordination* between speakers, both verbally and nonverbally, as speakers must pay attention to each other and take turns talking.

In VESSEL, we use these three characteristics of small talk to write an introductory small talk session for the exercise, wherein the coach discusses the topic of volunteer work with the learner. The session consists of a number of possible phrases and questions that the coach can say, ordered in a particular way to ensure a logical conversation flow; see Appendix B, which can be found on the IEEE Xplore Digital Library at http://ieeexplore.ieee.org/document/7915719/, for an overview of this. To establish familiarity, the coach

TABLE 2
VESSEL ECA Coach Affective Support Categories

Description	General	Specific	Very Specific
Reflective listening. This utterance makes explicit what emotional state the coach is perceiving, and (if applicable) the issue that's causing this state. This is put in the form of a statement, not a question. The learner has the chance to provide feedback if the coach's read is incorrect.	"It looks like you are experiencing difficulties."	"It looks like you are afraid."	"It looks like you are afraid of what could happen, if you fill out this form incorrectly."
Normalizing. This utterance puts the learner's issue and emotional reaction in a broader context, to show them that they are not alone.	"Many people encounter these difficulties."	"Many people become afraid in these circumstances."	"Many people become afraid in these circumstances."
Affirmation. This utterance tells the learner that the coach understands their emotional reaction, which is 'normal' (i.e., not exceptional or strange).The coach then helps the learner move look forward, by suggesting an action they can take, reminding them about help they can receive, or giving moral support.	"It is not strange that this is challenging for you. With practice, you will get better."	"It is not unusual for you to be afraid here. Keep trying, and you will see that it is not as difficult as you think."	"It is not unusual for you to be afraid of this. But this form is only a first step. In the interview afterwards, you will be able to clarify what volunteer work you do or do not want to do."
Self-efficacy supporting. This utterance tries to raise the learner's self-efficacy regarding the exercise topic and/or their skill in doing the exercise.	"I think, that you have already achieved a lot today."	"I think, that you have already achieved a lot today."	"I think, that you have already achieved a lot today."

Describes exact rules for creating utterances to match each support category, and includes example utterances used to provide support to a learner experiencing fear, the general, specific, and very specific affective support levels. Note that the utterances in the 'normalizing' and 'self-efficacy supporting' categories are very similar, while the utterances in the 'reflective listening' and 'affirmation' categories change strongly.

only discusses the established topic of small talk. To evoke solidarity, the coach both asks the user about their volunteer work experiences, and talks about their own 'experiences with volunteer work'. The coach asks follow-up questions whenever possible, but does not push learners if they are not interested in answering. To establish coordination, the coach follows a simple operation schema: whenever the learner starts talking after a question, the coach does not interrupt. Whenever the learner stops speaking, the coach waits three seconds, then utters the next phrase or question that makes sense in the scenario.

2.3 Technology: Emotion Measurement Tools

Emotion measurement tools can be grouped in three categories: psychological, physiological, and behavioural [35]. *Psychological* tools are subjective self-report tools, such as questionnaires. These tools are inexpensive, unobtrusive, and non-invasive, and they are the only way to measure a participant's inner perception and experience [36]. However, language barriers and cultural differences in emotion (see [21]) can make results unreliable. Participants may also be unwilling to talk about their emotional state to researchers, particularly in embarrassing cases, or they may be unable to put their emotional state to words. Finally, emotional self-reporting can be difficult in parallel with an experimental task without causing interference [36]. In practice, these tools can only be used after experimental sessions or in-between exercises.

Physiological tools are objective measures that use sensors. For instance, heart rate and galvanic skin response can be measured to determine arousal. Sensors can provide a continuous objective monitoring of the person's state [35] without being disruptive of task performance [37]. However, sensors can be invasive or intrusive, which could potentially influence the user's experience [36]. Sensors also often require specialized equipment and technical expertise to be used correctly. Using them sub-optimally, or in the presence of confounding circumstances such as excessive lighting or heat, may result in noisy data [35].

Lastly, *behavioural* tools measure motor-behavioural expressions and changes in physiological state. Unlike physiological tools, which are directly interested in the state of the body, behaviour tools measure body state in order to assess behaviour. Commonly, non-intrusive devices like computers and microphones are used: Zimmermann et al. [38] describe an example where "[the method] *extracts motor-behavioral parameters from log-files of mouse and keyboard actions, which can be used to analyze correlations with affective state*" (p. 540). The user's actions and behaviour can be used to predict and assign valence and arousal scores [38]. This approach is not very invasive (but Wong [36] notes that participants consider video cameras to be obtrusive), doesn't interfere with task performance, and can detect emotional cues that other tools cannot measure, such as facial expressions. However, special hardware and software are needed to capture this kind of data [39], and interpreting it requires trained, experienced and objective observers [35]. Additionally, the interpretation methods are commonly tested on 'produced' emotional expressions; in natural situations, recognition accuracy can drop harshly in case of spontaneous emotions [36].

In VESSEL, we combine all three tool types, to make use of the advantages of each. Questionnaires are administered at the start and end of each exercise to gauge users' self-reported affective state. We expect all participants to fill out these

questionnaires, making this a reliable source of data that will be useful for statistical evaluation of VESSEL's affective learning experience. Physiological and behavioural tools are used during exercises. The Shimmer sensor package [40] is used to measure learner arousal, using a photoplethysmographic (PPG) sensor attached to the earlobe (see Fig. 5). PPG sensors measure changes in light absorption that result from subcutaneous blood flow, which is translated into a measure of heart rate. The FaceReader facial recognition software package [41] uses a webcam to capture video of the learner's face, measure learner valence, and attempt to identify the occurrence of six basic emotions: happiness, anger, sadness, fear, surprise, and disgust (the latter two of which we are not interested in). Both body sensors and facial recognition let us rapidly assess users' affective states, and offer immediate affective support that is accurately tailored to that state; the FaceReader provides more detailed evaluation of the user's affective state, making it especially useful for support provision, while the Shimmer's arousal detection is a more objective measure of participant physiological state over time that may also be useful for later quantitative analysis.

3 Specification

3.1 Operationalization

We take three steps to operationalize the cognitive, affective, and social support behaviour for our prototype ECA coach. As a baseline for the prototype, we adopt the cognitive support model created in [11]. This model contains a systematic way to create cognitive support utterances to cover all potential cognitive difficulties in an exercise, a comprehensive set of behaviour rules for an ECA coach to provide cognitive support, and rules to model speech recognition. Cognitive support utterances are created for all identified 'difficult elements' in the exercise, for each of [11]'s five levels of cognitive support. We incorporate the rules for when and how to provide cognitive support into this prototype as they are. Since cognitive support is not evaluated with this prototype, the full process is left out of this paper (but interested readers are referred to [11]). Similarly, we apply the techniques for emulating speech recognition: we create a dictionary of keywords (based on exercise 'difficult elements' and cognitive support categories), and define how the coach can react to these keywords. We make one significant change: during the recruiter part of the exercise, cognitive support utterances will be spoken by the recruiter ECA, not the coach ECA. This change is made because our earlier work with two concurrent ECAs [4] shows that asking a coach character for help in a conversation exercise interrupts the dialogue, and leads to learner confusion.

To operationalize affective support, we use the four motivational interviewing categories and three specificity levels from Table 2 to create affective support utterances: for every emotion the coach ECA can recognize except happiness (which does not require support), we create one or two utterances for every category-level combination. We then define when and how these utterances are used. We say that affective support must be given in three circumstances. If learner arousal is high but valence is not clearly low, or if valence is low but arousal is not clearly high, the coach detects that affective issues are happening, but cannot quantify which. In this case, General support is given (as per Table 2). If arousal and valence clearly indicate anger, fear, or sadness as per Table 1, or if the FaceReader program strongly detects anger, fear, or sadness, the coach gives affective support on the Specific level, tailored to that emotion. If the cause of the anger, fear, or sadness is also clearly detected, the coach gives affective support on the Very Specific level, tailored to that emotion and cause. We have identified a number of elements and situations in the exercise that will likely lead to particular affective reactions, for which Very Specific affective support utterances were recorded. For example, when the recruiter ECA very curtly asks questions, we suspect low-literate participants will get angry at this disrespectful style of speaking. When giving affective support, the coach uses four utterances in sequence: reflective listening, normalization, affirmation, and self-efficacy supporting. All utterances are 5 seconds apart. After giving all four types of affective support utterance in a row, the coach must wait at least one minute before giving more: this is done to prevent endless repetition of the same support for learners that stay in the same affective state for a longer time.

Finally, we operationalize social support by using the small talk utterance corpus in Appendix B, available in the online supplemental material. The coach uses the utterances as indicated. We define the speech recognition options for small talk here: when the coach asks the learner a question, it can understand all varieties of 'yes' or 'no' as answers, and react accordingly. As long as a learner is talking (to answer a question, or for other reasons), the coach recognizes this and does not talk or interrupt. When the learner is not talking, the coach moves through small talk utterances, keeping 5 seconds between each.

3.2 Requirements Baseline

Section 2's foundation data are now used to refine the existing VESSEL requirements baseline. Only those requirements that change on the basis of the expanded foundation are refined, for the *coach* and *exercises* aspects of VESSEL; requirements that are not described do not change. Table 5 (see Appendix A, available in the online supplemental material) shows the refined requirements baseline.

Requirement *R1. Adaptability* is refined for both coach and exercises. The coach (**R1.1-C**) should ensure that affective support matches the learner's emotional state. Affective support must only be given if the sensors indicate particular emotional valence and intensity. The exercises (**R1.1-E**) should be cognitively and affectively challenging. An exercise is affectively challenging if learners experience significant anger, fear, or sadness at least once while doing the exercise.

Requirement *R2. Sensitivity* is refined for the exercises (**R2.1-E**), as an extension of **R1.1-E**. Exercises must be as sensitive *or insensitive* as needed to reach intended difficulty levels. Specifically, in exercises that feature conversation partners, the conversation partner's dialogue must display the right level of sensitivity to effect the intended affective difficulty. If the conversation partner is too kind, no affective difficulty is reached (see [4] for an example of this), but if the conversation partner is too abrasive, low-literate users might stop the exercise midway (see [1]).

Requirement *R6. Support* is zoomed in to coach-offered affective support (**R6.2-C**) and social support (**R6.3-C**). The

Fig. 4. The two appearance options for the ECA coach.

coach should offer affective support and social support according to the behaviour rules in Section 3.1. Social support should be offered before the exercises, and affective support should be offered during the exercises, concurrent with cognitive support (**R6.1-C**, remaining unchanged). No support is offered after exercises.

Requirement *R7. Interactivity* is refined for the coach and the exercises. The coach's proactive affective support interactions with the learners (**R7.1-C**) should be driven by sensors and facial recognition, and all proactive support interactions should be guided by the rules in Section 3.1. If (during an exercise) cognitive learning support is offered in an exercise that has a conversation partner present (*such as the recruiter ECA*), the utterance should be spoken by the conversation partner ECA instead of the coach ECA (**R7.1-E**). This applies to both proactive and reactive cognitive support.

4 Evaluation: Prototype Development

Functionality. The prototype consists of the two-part volunteer work exercise described in Section 2.1, and an ECA coach that offers cognitive, affective, and social learning support according to the rules and timing approaching described in Section 3.1.

Interaction Methods. Learners interact with the form part of the exercise using mouse and keyboard. They can talk to the coach and the recruiter in natural language. For the purposes of evaluation, coach and recruiter are designed to be controlled via the Wizard-of-Oz method [10] similar to [11], meaning that speech recognition is emulated by the Wizard operators. This must be done in accordance with the speech recognition rules in Section 3.1.

Appearance. The coach ECA avatars were developed in Unity. Two visual variations of the coach were made to match the two variant exercises (Section 2.1). Both were based on the appearance of our previous coaches [4], [11]. Fig. 4 shows the two coach appearances. The recruiter ECA avatars (Fig. 2) were also created in Unity.

5 Evaluation: Methods

5.1 Experimental Design

An experiment was designed to evaluate the learning effectiveness of our VESSEL prototype. We wanted to compare this prototype, which offers 'full' (cognitive, affective, and social) learning support, to a prototype that offers only cognitive learning support, built according to our previous specification [11]. Six hypotheses were created, corresponding to the six learning effectiveness outlined earlier: cognitive, affective, and social learning experience, and cognitive, affective, and social learning outcomes. In general, we expect that the current 'Full Support' prototype results in higher learning effectiveness on all fronts than the 'Cognitive Support' prototype.

Fig. 5. Shimmer sensor with PPG ear clip.

H1. *Cognitive Experience (Performance).* Learners that receive full learning support report better performance during the exercise, expend less effort doing the exercise, complete the exercise quicker, and receive more support while doing the exercise, than learners that receive only cognitive learning support.

H2. *Affective Experience (Positive Affect).* Learners that receive full learning support report a more positive affective state than learners that receive only cognitive learning support.

H3. *Social Experience (Motivation).* Learners that receive full learning support are more motivated to learn and to continue learning than learners that receive only cognitive learning support.

H4. *Cognitive Outcomes (Success).* Learners that receive full learning support remember and recall more details about the exercise than learners that receive only cognitive learning support.

H5. *Affective Outcomes (Self-Efficacy).* Learners that receive full learning support report a higher increase in self-efficacy than learners that receive only cognitive learning support.

H6. *Social Outcomes (Coach Opinion).* Learners that receive full learning support have a more positive view about the coach, and initiate more interactions with the coach, than learners that receive only cognitive learning support.

To evaluate these hypotheses, we designed a mixed-method repeated-measures experiment that combined within-subjects and between-subjects measurements. The main independent variable was *Support Model*, with two levels: *Full Model*, and *Cognitive Model*. Participants were invited to work with both prototypes: in one experimental session, participants completed the exercise twice, once with the full prototype and once with the cognitive prototype. Prototype order was counterbalanced: 50 percent of participants did the Full Model condition first and the Cognitive Model condition

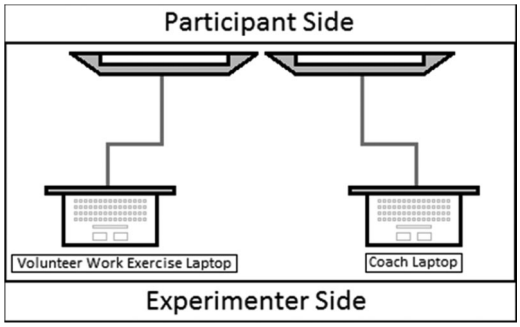


Fig. 6. Schematic overview of experimental setup. Two laptops (bottom figures) connect to two monitors (top figures). Located near the monitors are also: a keyboard, a mouse, a microphone, and a webcam (not shown in image).

second, and 50 percent did the opposite. The two versions of the coach and the two exercises were counterbalanced as well, leading to eight different orders.

5.2 Measures

Seventeen quantitative dependent variables were measured. Twelve were self-report questions, measured using three questionnaires (Section 5.4), and five were objective performance metrics. Appendix D, available in the online supplemental material, shows the variables.

5.3 Participants

Participants were recruited from five reading and writing classes throughout the Netherlands (located in Rotterdam, Nijmegen, and Den Helder). We used [3]'s five language learner profiles to select participants for this study. Only learners that matched profiles 2, 3, and 4 were invited: learners in profiles 1 and 5 are respectively too skilled to benefit from our level of exercise and support, and too low-skilled to independently engage with the language level and complexity level of our prototype. Thirty-four participants completed the entire experiment: twenty men and fourteen women, with ages ranging from 19 to 64 ($M = 41.3$, $SD = 15.1$). Ten participants self-identified as being natively fluent in Dutch. The other twenty-four identified as 'somewhat fluent'. Other languages spoken, either natively or as a second language, included: Arabic, Amharic, Aramaic, Bosnian, Catalan, Dari, Edo, English, Farsi, French, Italian, Moroccan, Papiamentu, Russian, Somali, Spanish, Swahili, Swedish, Tamil, and Turkish. Twenty participants reported having prior experience with volunteer work.

5.4 Materials

The experimental setup consisted of two laptops, each connected to one monitor (Fig. 6), which were used to run the experiment. The monitors allowed participants to see and interact with the exercises. The laptops allowed experimenters to run the exercise (left laptop) and the coach (right laptop). On the participant side, a mouse, keyboard, speakers, and microphone were provided: mouse and keyboard let

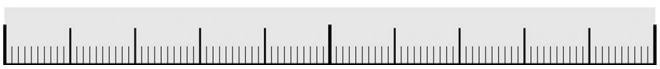


Fig. 7. Visual analogue scale used in the PAQ and ERQ.

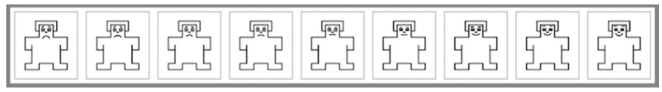


Fig. 8 . Self-Assessment Manikin bar used to measure PAQ.7.

participants fill out the form, speakers played the coach and recruiter ECA utterances, and the microphone was used to suggest that participants could talk to the ECA characters, as well as to record audio (with consent). A webcam was attached to the left monitor, to capture visual data for the FaceReader software: participants were told this allowed the coach to 'see' them. One Shimmer sensor was used (Fig. 5).

Three questionnaires were used. Two questionnaires measured the fifteen self-report variables shown in Appendix D, available in the online supplemental material. The 'participant assessment questionnaire' (PAQ) measured participant self-efficacy on four topics (reading Dutch, computer use, filling out a form, and having a conversation in Dutch), participant motivation to learn, participant fear of going to school, and participant affective state on the dimensions of valence, arousal, and dominance. The 'exercise reflection questionnaire' (ERQ) measured participants' view on their performance and exercise results, as well as their view of the coach. Two answer methods were used. Answers to the three participant affective state questions (PAQ.7, PAQ.8, and PAQ.9 in Appendix D, available in the online supplemental material) were given using the Self-Assessment Manikin (Fig. 8, see [27]). Answers to all other questions were given using a visual analogue scale (Fig. 7). Questions were read aloud to participants, who would then mark answers on the corresponding bar; this method ensures that participant reading and writing skills are not a factor in accurate answering. The fourth 'demographic' questionnaire measured: participant age, sex, time period lived in the Netherlands, known languages, and prior volunteer work experience.

For objective measures, exercise completion time was measured using a digital clock and a stopwatch. The number of coach support utterances received by participants was recorded by hand, and categorized in the following way: utterances were either cognitive or affective support utterances (social support utterances were not recorded), they were recorded during the form part or the recruiter part of the exercise, and they were initiated either by the coach or by the participant. Finally, a 'recall test' was created to measure learning success. After each exercise, participants were given one minute to name as many form elements as they could remember. Researchers wrote down which of the five categories on the form (see Fig. 3) participants named. Score was calculated per category: 1 point if the category was named and described correctly, or 0.5 points if either the category was named correctly, but not described, or if it was described (ex. by giving examples of category contents) but not named, up to a maximum score of 5 points.

Note here that discrete participant emotional states as measured by the Shimmer and FaceReader (happiness, anger, sadness, and fear) are not used for hypothesis evaluation. As described in Section 2.2.1, this basic model of emotions is useful for driving immediate coach decisions.

We instead use the Self-Assessment Manikin (Fig. 8), which is an expression of the dimensional approach, as we believe this allows for more in-depth analysis of emotional states.

5.5 Procedure

Each thirty-minute session started with a general introduction, informed consent forms, and the demographic questionnaire. Researchers explained the general experiment flow and the experimental setup hardware (Fig. 6). The Shimmer sensor was introduced and attached to the participant's earlobe: doing this before measurements were taken gave researchers time to calibrate the exact placement for optimal results, and allowed participants to get used to the sensation. The first PAQ was administered. Researchers then activated the designated coach prototype. The ECA coach (controlled Wizard-of-Oz style by one researcher who followed the control rules described in Section 3.1) introduced itself to the user and explained the first exercise. Participants were told to complete the exercise in two steps: fill out the form, then have a conversation with the recruiter. Participants were given as much time as needed to complete the exercise. After the first exercise, the first ERQ and the second PAQ were administered, followed by a first recall test. Researchers then activated the other coach prototype, after which the second exercise was introduced and conducted similar to the first. After the second exercise, a second ERQ and third PAQ were administered, and a second recall test. Finally, participants were fully debriefed (including a behind-the-scenes look of the VESSEL prototype and the Wizard-of-Of method) and rewarded for participation.

6 Evaluation: Results

Four analysis steps are presented here. Section 6.1 shows the evaluation of the six main hypotheses (Section 5.1). Section 6.2 looks at potential order and learning effects between the two exercises. Section 6.3 describes qualitative observations made by the researchers, during the experiment and by listening to recorded audio proceedings afterwards. Based on observations and initial results, Section 6.4 shows two post-hoc analyses. Before analysis, the data was characterized and checked for irregularities. No obvious mistakes or irregularities were found. Questionnaire reliabilities were assessed: Cronbach's alpha was .773 for the PAQ (.810 if based on standardized items) and .872 for the ERQ (.844 for standardized items). No data reduction measures were used.

6.1 Hypothesis Evaluation

To evaluate the six hypotheses data from the PAQ, ERQ, and DM data were subjected to repeated measures General Linear Model (GLM) analyses. The ERQ and DM data were analyzed with one two-level factor: *Coach Type*, with levels 'Full Coach' and 'Cognitive Coach', The PAQ data were analyzed with one three-level factor: *Coach Type*, with levels 'Before Exercise', 'Full Coach' and 'Cognitive Coach'. Three significant results were found. In the Full Coach condition, participants received more affective support than participants in the Cognitive Coach condition, in both form ($F = 14.431,\ p = .001,\ \beta = .957$) and recruiter ($F = 52.755,\ p = .000,\ \beta = 1.000$) parts, as well as more 'total' support (cognitive and affective support combined) during the form part ($F = 29.005,\ p = .000,\ \beta = .999$). This supports H1: learners receive more support during the form part of the exercise. Also, participants in the Full Coach condition actively initiated more learner-coach interactions, i.e., asked the coach more questions without prompting, than participants in the Cognitive Coach condition ($F = 8.484.,\ p = .007,\ \beta = .806$). This supports H6: learners engage in more self-started interaction with the coach. Appendix C, available in the online supplemental material, shows the full results.

6.2 Exercise Order Effects

To test for exercise order effects, the PAQ, ERQ, and DM data were used in two more repeated measures GLM analyses. Similar to the previous, the ERQ and DM data were analyzed with two-level factor (*Exercise Order*, with levels 'First Exercise' and 'Second Exercise') and the PAQ data were analyzed with one three-level factor (*Exercise Order*, with levels 'Before Exercise', 'First Exercise' and 'Second Exercise'). The following significant results were found. Self-efficacy about reading Dutch ($F = 3.848,\ p = .032,\ \beta = .562$) and about volunteer work ($F = 5.635,\ p = .008,\ \beta = .825$) were higher after the second exercise than after the first exercise. Participants also reported lower arousal after the second exercise ($F = 4.754,\ p = .036,\ \beta = .562$). Related to the form part of the exercise, completion time ($F = 16.042,\ p = .000,\ \beta = .972$), number of cognitive support utterances received ($F = 8.403,\ p = .007,\ \beta = .802$), and total amount of support utterances received ($F = 5.049,\ p = .032,\ \beta = .586$) were all lower after the second exercise. Finally, related to the recruiter part of the exercise, participants started more learner-coach interactions during the second exercise ($F = 9.782,\ p = .004,\ \beta = .858$). Appendix C, available in the online supplemental material, shows the full results.

6.3 Observations

Experimental observations showed that in general, learners managed to use the prototype and work with the coach as intended. Learners engaged with and completed the exercises, and they listened to and asked for help from the coach. This was particularly true for the Cognitive Support prototype, which was observed to work almost exactly like the prototype in our previous experiment (see [11]). The provided cognitive support was sufficient to help learners in both the form and recruiter parts of the exercise. As in [11], learners almost always listened to coach advice, and would only occasionally ask questions themselves. In the recruiter part of the exercise, participants adapted to talking/listening to the recruiter ECA with no problems, and had no problems with the recruiter providing cognitive support. One unexpected side effect was that participants would almost 'forget about the coach', since in this prototype it had no other support to give. But this did not seem to negatively influence the relation between learner and coach, with learners mostly expressing amusement, "*Oh, she's still here too!*", whenever they noticed the coach (still visible on the right screen).

In the Full Support prototype conditions, the 'small talk' social support worked almost exactly like the small talk in

our first proof-of-concept prototype [4]. Participants seemed to honestly and genuinely speak with the coach about their volunteer work experiences and preferences. Cognitive support in this condition worked similar to the Cognitive Support condition, the only difference being that researchers felt that participants actively asked more questions. However, the affective support seemed to work only piecemeal. Providing affective support was hampered because our sensors did not work as well as hoped. The Shimmer data was noisy, and prone to halting and resuming at random moments. The FaceReader data generally gave clearer reads on participants' emotional states. However, we encountered the issue that some participants had resting facial expressions that the FaceReader interpreted as a particular emotion: for instance, one participant's facial features were interpreted as a high level of 'sadness' all the time, leading to the coach repeating similar affective support every minute. Another problem was that both Shimmer and FaceReader had serious difficulty working with darker skin tones: the FaceReader algorithm was less effective at reading black and tan faces, and the Shimmer's PPG (which works by sending red light through the earlobe) seems to have been calibrated on light skin, not taking the different light absorption/reflection profile of dark skin into account. As a result, in the cases of many dark-skinned participants (who made up a significant subset of the NT2 group), we simply did not have enough accurate data to provide affective support to begin with. As a solution, we decided to incorporate the personal situational interpretation of the wizard operators into the decision making process: if both researchers agreed that a participant was clearly exhibiting a certain emotion, affective support could be provided. In practice, this agreement was not reached very often, and as a result, the provision of affective support to these participants was limited.

In practice, the Very Specific level of support was never used, as the situations we expected to necessitate this support (and recorded these utterances for) were not seen. The reception of the General and Specific support levels was mixed. Participants did often verbally acknowledge the support (for instance, by responding to or thanking the coach). And during debriefing, participants would often mention that the Full Support coach "*cared about* [the participant] *more*". However, sensors never showed a direct physiological reaction to affective support (that was clear enough to discern with accuracy). This makes it unclear to what degree the affective support had the intended effects. One other unexpected observation was that some participants would countermand the coach's affective support: reflective listening statements like "*It looks like you are scared*" were sometimes met with negations such as "*No I'm not*." We added to our control rules that, in these cases, the rest of the affective support for this occurrence should be cancelled.

Finally, unexpected differences were seen between the NT1 and NT2 participant groups. The NT1 participants were generally better at the recruiter part of the exercise, due to their native Dutch speaking and large vocabulary, but worse during the form part of the exercise, due to limited ICT and computer skills. The NT2 participants showed the inverse: good computer skills, but limited Dutch vocabulary. While these differences have been seen to some degree in our earlier work (see particularly [1] for our overview of meaningful differences), this experiment marks the first time in our evaluation of VESSELprototypes that significantly different outcomes were found between the groups (as per Section 6.3).

6.4 Post-Hoc

Based on the aforementioned observations and analysis results, we decided to investigate the effect of learner background, or 'type'. Two types of learners were identified: 9 participants were learners with a native Dutch background ('NT1'), and 25 participants were learners with a migrant background ('NT2'). The previous repeated measures GLM analyses were then repeated, using *Learner Type* as a between-subjects variable. Results suggest significant differences between the experiences of the two types. NT1 learners reported higher (self-reported) performance ($F = 4.585$, $p = .040$, $\beta = .547$), higher valence ($F = 5.918$, $p = .021$, $\beta = .655$), less received cognitive support in both the form ($F = 4.586$, $p = .040$, $\beta = .545$) and conversation ($F = 5.350$, $p = .028$, $\beta = .610$) parts, and lower completion time in the recruiter part ($F = 10.387$, $p = .004$, $\beta = .871$). NT2 learners reported higher computer use self-efficacy ($F = 4.171$, $p = .025$, $\boldsymbol{\beta} = .692$), more self-initiated coach interaction in the recruiter part ($F = 8.589$, $p = .004$, $\beta = .850$), and a higher desire to use the coach again in the future ($F = 7.508$, $p = .010$, $\beta = .757$). Additionally, we found one interaction effect for Learner Type and Exercise Order: NT1 learners reported spending high effort on the first exercise and low effort on the second, while NT2 learners reported moderate effort on the first exercise and high effort for the second ($F = 9.888$, $p = 004.$, $\beta = .862$).

We also tested the variables of age, sex, time spent living in the Netherlands, experience with volunteer work, and counterbalancing order for between-subjects effects. Three significant effects were found for *Learner Sex*. Women received more affective support than men, overall ($F = 9.333$, $p = .005$, $\beta = .840$). An interaction effect between Learner Sex and Coach Type showed that during the form exercises, men received a higher number of support utterances in the Cognitive Coach condition, and women received a higher number of support utterances in the Full Coach condition ($F = 6.049$, $p = .020$, $\beta = .663$). Finally, an interaction effect between Learner Sex and Exercise Order showed that for women, self-efficacy with regard to holding a conversation was significantly higher after Exercise 1 than either at the start of the experiment, or after Exercise 2. For men, this difference did not exist ($F = 3.586$, $p = .040$, $\beta = .621$). Appendix C, available in the online supplemental material, shows the full results of both Learner Type and Learner Sex analyses.

7 Conclusions

This study aimed to answer two research questions. Question Q1 was: "*How can we create a design specification for VESSEL that incorporates rules for cognitive, affective, and social learning support provided by an ECA coach?*" This question was answered in sections 2 through 4. Sub-question Q1a, "*Which emotional models, motivational interviewing rules, small talk scenarios, and measurement methods are needed to create*

these rules?", was answered in Section 2. An overview was created of operational demands (a description of the volunteer work exercise to be used in the prototype), human factors knowledge (the three kinds of extant emotional models, and our systematic interpretations of motivational interviewing and small talk), and technology (three kinds of autonomous emotion measurement tools). This overview was incorporated into the sCE foundation of our VESSEL design specification: we created the volunteer work exercise, made rules to describe motivational interviewing and small talk behaviour, and selected the Shimmer and FaceReader sensors for use with the prototype. Sub-question Q1b, "*Which functionalities, interaction methods, and appearances should the ECA coach have to reflect this specification?*", was answered in sections 3, where the design specification was updated with new control rules and functional requirements, and 4, where a new VESSEL prototype was created based on the updated specification.

Question Q2 was: "*Does an ECA coach created in accordance with this specification result in a higher learning effectiveness for low-literate learners than an ECA coach that incorporates only formalized cognitive learning support?*" This question was answered in Sections 5 and 6, where we experimentally evaluated the prototype by comparing it against a prototype built according to our previous design specification [11]. Six hypotheses were tested:

- H1. *Cognitive Experience (Performance).* This hypothesis is partially supported. Learners in the Full Support condition did receive significantly more learning support than learners in the Cognitive Support condition, but did not report better performance, expend less effort, or complete the exercise quicker.
- H2. *Affective Experience (Positive Affect).* This hypothesis is not supported. There was no significant difference in affective state during and after the exercises between learners in either condition.
- H3. *Social Experience (Motivation).* This hypothesis is not supported. There was no significant difference in motivation to learn and to continue learning between learners in either condition.
- H4. *Cognitive Outcomes (Success).* This hypothesis is not supported. There was no significant difference in recall test results between learners in either condition.
- H5. *Affective Outcomes (Self-Efficacy).* This hypothesis is not supported. There was no significant difference in self-efficacy increase between learners in either condition.
- H6. *Social Outcomes (Coach Opinion).* This hypothesis is partially supported. Learners in the Full Support condition initiated more interactions with the coach than learners in the Cognitive Support condition, but they did not report a more positive view about the coach.

These hypothesis results seem to indicate that few differences exist between the Full Support and Cognitive Support coaches. Of the two partially supported hypotheses, H1 does not provide much information: that learners in the Full Support condition receive more support overall is easily explained by the fact that these learners received cognitive *and* affective support during the exercise, where learners in the Cognitive Support condition *only* received cognitive support. Hypothesis H6 does show an interesting finding: learners in the Full Support condition were quicker to proactively talk to the coach. This finding matches our expectation that adding affective and social support makes it clearer to learners that the coach can be talked to like a human conversation partner. In our previous work, learners that used the 'proof-of-concept' prototype (which included early operationalization of affective and social support, see [4]) were observed to proactively speak with the coach more than learners that used our first cognitive support prototype [11]. The results in this study now statistically validate these observations. It is currently unsure what mechanisms lead to this increased 'affordance of being spoken to'. Future studies could try to disentangle the effects of affective and social support, to see if either can be pinpointed as the cause, or if the effect only happens with a combination of support types.

Based on the results presented above, we must conclude that very little differences existed between the Full Support and Cognitive Support prototypes. The addition of affective and social learning support did not have many of the predicted effects. Three potential explanations are offered here. The first explanation is that our affective support manipulations may not have been large enough to produce an effect. Section 6.4 describes how sensor problems led to issues with the provision of affective support. In practice, the researchers only confidently employed affective support in a limited number of situations. This should be considered an oversight on our part: We expect that better results can be obtained by extensively testing and calibrating the Shimmer and FaceReader sensors to our participants. The cognitive appraisal method (described in Section 2.2.1) to do could possibly be used for this. Since emotion sensing technology is still showing shortcomings in *in situ* applications like this one, future work should investigate whether or not incorporating this method in the design of affective support is valuable. Also, the volunteer work scenario used in the prototype did not seem to result in a great deal of affective challenge. Selecting an exercise for this goal is difficult: while some crucial practical situations have obvious affective impact (such as health-related issues like hospitalization, or death of a family member) it was considered ethically unjustifiable to use this level of affective stress to evaluate a digital coach. The volunteer work scenario was seen as having the potential to be affectively challenging, which we could bring out by carefully designing the difficulty and the affective and social behaviour of the recruiter ECA (see Section 2.1). It is unclear whether or not this worked. The second explanation is that our approach to providing affective learning support with an ECA has been too limited. Studies indicate that ECAs in general can potentially change the affective experience of doing computer exercises [42] and emotionally connect to learners [43]. The *Multimodal Affective and Reactive Character* framework [44] describes three factors that influence the effectiveness of affective characters: The capacity of the agent to respond to the user in real-time, subtle visual indicators of agent affective state, and the ability to express differences in affective reactions to different individual learners. VESSEL can act in real-time and adapt to individual learners, but does not use

visual indicators of affective states: The appearances and facial expressions of our coach and recruiter ECAs were more or less static (see Figs. 3-4). The way embodied characters look impacts how their functionality and possibilities are perceived (see [45]). It is possible that the more 'stylized' (non-realistic and exaggerated) appearance of our coach (Fig. 4) impacted this, and that a more 'naturalistic' (human-looking) appearance, including affectively expressive facial expressions, would have served us better [44], [46]. The third explanation is that our experimental setup impacted coach effectiveness. Participants used both coach versions in a span of 30 minutes: this may have caused them to see both coaches as a single entity, with slightly different behaviour between exercises. Support for this is offered by a debriefing observation: when participants were asked if they noticed any differences between the two coach ECAs, many would say yes, and then describe differences between the coach ECAs and the recruiter ECAs. In this scenario it is possible that the coach did have affective effects, but that the attribution of these effects (particularly in questionnaires) was confounded by the presence of the recruiter. Future studies should try to disentangle these effects more clearly: Maybe exercises with conversation components should use the coach as the conversation partner directly, or maybe the coach ECA should be hidden from view in these cases. In general, the lack of significant results for our affective learning support is counter to expectations (see [42], [43], [44], [47]), and future work in this direction should focus on investigating ways to resolve this.

Results for the order effects evaluation (Section 6.2) show that the prototype in general did work as expected. Participants always completed all exercises using cognitive support, similar to our previous cognitive support prototype [11]. Participants accepted and understood the coach, and used its help to get through the exercises when needed. The lower completion time and lower need for learning support in the second exercise compared to the first exercise indicates a straightforward (and expected) learning effect. The lower arousal and the increased number of self-initiated learner-coach interactions in the second exercise seems to suggest that learners were more 'at ease' with the system the second time around. Finally, the increase in self-efficacy (with regard to 'reading Dutch' and 'volunteer work') over the exercises is interesting, as this reproduces our findings in [4] and [11]. Appendix C, available in the online supplemental material, shows clearly that the self-efficacy increase happened after the first exercise. Learners judge their self-efficacy lower before doing any exercise, judge it higher after completing an exercise for the first time, and then stay on that higher level throughout. One of the strongest sources of self-efficacy information is successfully completing a task yourself ('enactive mastery') [48]. Results from all our prototype experiments suggest that working with VESSEL provides this: self-efficacy about a larger domain ('reading Dutch' or 'doing online banking') increases after completing a specific scenario exercise, and stays high. It would be interesting for future studies to investigate how long these self-efficacy increases last. For instance, does self-efficacy remain high after four or five exercises? And does self-efficacy remain high if longer amounts of time (i.e., weeks or months) pass between exercises?

Finally, the differences we found between NT1 and NT2 learners highlight the importance and added value of personalization. One possible explanation is that the volunteer work scenario in this prototype has caused this. This exercise is more grounded in Dutch society than online banking exercises [4], [11], and it is not a topic that both NT1 and NT2 learners have a lot of direct experience with (unlike [4]'s city hall passport exercise). In this prototype's exercise, NT1 and NT2 learners encountered different problems, and showed different reactions. Combined with the aforementioned sensor difficulties (for NT2 students), it is not surprising to find significant differences in outcomes for the two groups. The specific differences were not unexpected: that NT1 learners have better vocabulary and poorer computer skills than NT2 learners is entirely in line with literature expectations (see [1], and [3], which we used for participant selection in Section 5.3). The findings reinforce once more that 'low-literate learners' are a very heterogeneous group [49], and that learning for these learners must be personalized to their wants and needs to be effective [2]. Furthermore, the found differences between male and female learners indicate that personalization can be valuable for many attributes. From earlier work results, we see particular participant attributes as 'less important': age, sex, and schooling history did not show up as significant between-subjects factors in [4] or [11]. The fact that learner sex is now a significant factor here indicates that specific scenario, types of learning support, and other factors of learning context can play a major role in determining what kinds of personalization are valuable. These results are comparable to [47], who used an affectively expressive virtual storyteller character with children aged 6-10: The affective manipulation in that work had little main effects, but unexpected main and interaction effects showed difference between groups of children younger than 8 years and children aged 8 and above. Future work in this field should investigate along which personal attributes affective learning support for people of low literacy can best be personalized, using the results presented here as a starting-off point.

Acknowledgments

The authors would like to thank Knup Fuhri and Mirella Verspiek for their valuable contributions. This publication was supported by the 'Interaction For Universal Access' track of the Dutch national program COMMIT.

References

[1] D. G. M. Schouten, R. T. Paulissen, M. Hanekamp, A. Groot, M. A. Neerincx, and A. H. M. Cremers, "Low-literates' support needs for societal participation learning: Analysis methods, refinement models, and outcomes." in development.

[2] D. G. M. Schouten, N. J. J. M. Smets, M. Driessen, K. Fuhri, M. A. Neerincx, and A. H. M. Cremers, "Requirements for a virtual environment to support the social participation education of low-literates," *Univers. Access Inf. Soc.*, pp. 1–18, 2016.

[3] J. Kurvers, K. Dalderop, and W. Stockmann, "Cursistprofielen Laaggeletterdheid NT1 & NT2," Tilburg, 2013.

[4] A. A. Deneka, "A Coach ECA to increase societal participation of low literate and non-native citizens in the societal participation learning support system," University of Twente, Enschede, Netherlands, 2014.

[5] M. A. Neerincx, "Situated cognitive engineering for crew support in space," *Pers. Ubiquitous Comput.*, vol. 15, no. 5, pp. 445–456, 2011.
[6] M. A. Neerincx and J. Lindenberg, "Situated cognitive engineering for complex task environments," in *Naturalistic Decision Making and Macrocognition*, J. M. Schraagen, L. G. Militello, T. Ormerod, and R. Lipshitz, Eds. Aldershot, U.K.: Ashgate Publishing Limited, 2008, pp. 373–390.
[7] J. van de Pol and E. Elbers, "Scaffolding student learning: A micro-analysis of teacher–student interaction," *Learn. Cult. Soc. Interact.*, vol. 2, no. 1, pp. 32–41, Mar. 2013.
[8] W. R. Miller and S. Rollnick, "Ten things that motivational interviewing is not," *Behav. Cogn. Psychother.*, vol. 37, no. 2, pp. 129–140, 2009.
[9] J. Cassell and T. W. Bickmore, "Negotiated collusion: Modeling social language and its relationship effects in intelligent agents," *User Model. User-adapt. Interact.*, no. 13, pp. 89–132, 2003.
[10] D. Maulsby, S. Greenberg, and R. Mander, "Prototyping an intelligent agent through Wizard of Oz," in *Proc. ACM SIGCHI Conf. Human Factors Comput. Syst.*, 1993, pp. 277–284.
[11] P. Massink, "An intelligent tutoring system for low-literates to encourage societal participation," Utrecht University, Utrecht, Netherlands, 2015.
[12] J. van de Pol, M. Volman, and J. Beishuizen, "Scaffolding in teacher–student interaction: A decade of research," *Educ. Psychol. Rev.*, vol. 22, no. 3, pp. 271–296, Apr. 2010.
[13] D. Grandjean, D. Sander, and K. R. Scherer, "Conscious emotional experience emerges as a function of multilevel, appraisal-driven response synchronization," *Conscious. Cogn.*, vol. 17, no. 2, pp. 484–495, 2008.
[14] C. Darwin, P. Ekman, and P. Prodger, *The Expression of the Emotions in Man and Animals*. Cary, NC, USA: Oxford University Press, 1998.
[15] P. Ekman, "An argument for basic emotions," *Cogn. Emot.*, vol. 6, no. 3–4, pp. 169–200, 1992.
[16] R. Plutchik, *The Emotions*. Lanham, MD, USA: University Press of America, 1991.
[17] W. G. Parrott, *Emotions in Social Psychology: Essential Readings*. Park Drive, U.K.: Psychology Press, 2001.
[18] S. Schachter and J. E. Singer, "Cognitive, social and physiological determinants of emotional state," *Psychol. Rev.*, vol. 69, pp. 379–399, 1962.
[19] W. James, *The Principles of Psychology*, vol. 1. Mineola, NY, USA: Dover Publications, 1950.
[20] P. Ekman and W. V Friesen, "Measuring facial movement," *Environ. Psychol. Nonverbal Behav.*, vol. 1, no. 1, pp. 56–75, 1976.
[21] P. C. Ellsworth and K. R. Scherer, "Appraisal processes in emotion," *Handb. Affect. Sci.*, p. 572, 2003.
[22] M. A. Nicolau, H. Gunes, and M. Pantic, "Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space," *IEEE Trans. Affect. Comput.*, vol. 2, no. 2, pp. 92–105, Apr.-Jun. 2011.
[23] J. A. Russel, "A circumplex model of affect," *J. Pers. Soc. Psychol.*, vol. 39, no. 6, pp. 1161–1178, 1980.
[24] A. M. Oliveira, M. P. Teixeira, I. B. Fonseca, and M. Oliveira, "Joint model-parameter validation of self-estimates of valence and aurosal: Probing a differential-weighting model of affective intensity," *Proc. Fechner Day*, vol. 22, no. 1, pp. 245–250, 2006.
[25] P. A. Lewis, H. D. Critchley, P. Rotshtein, and R. J. Dolan, "Neural correlates of processing valence and arousal in affective words," *Cereb. Cortex*, vol. 17, no. 3, pp. 742–748, 2007.
[26] J. A. Russel and A. Mehrabian, "Evidence for a three-factor theory of emotions," *J. Res. Pers.*, vol. 11, no. 3, pp. 273–294, 1977.
[27] M. M. Bradley and P. J. Lang, "Measuring emotion: The self-assessment manikin and the semantic differential," *J. Behav. Ther. Exp. Psychiatry*, vol. 25, no. 1, pp. 49–59, 1994.
[28] S. Rollnick and W. R. Miller, "What is motivational interviewing?," *Behav. Cogn. Psychother.*, vol. 23, no. 4, pp. 325–334, 1995.
[29] W. R. Miller, "Motivational interviewing with problem drinkers," *Behav. Psychother.*, vol. 11, no. 2, pp. 147–172, 1983.
[30] S. A. H. Friederichs, A. Oenema, C. Bolman, J. Guyaux, H. M. van Keulen, and L. Lechner, "I Move: Systematic development of a web-based computer tailored physical activity intervention, based on motivational interviewing and self-determination theory," *BMC Public Health*, vol. 14, no. 1, 2014, Art. no. 212.
[31] L. C. Sobell and M. B. Sobell, "Motivational interviewing strategies and techniques: Rationales and examples," 2008, http://www.nova.edu/gsc/forms/mitechniquesskills.pdf
[32] K. M. Emmons and S. Rollnick, "Motivational interviewing in health care settings: Opportunities and limitations," *Am. J. Prev. Med.*, vol. 20, no. 1, pp. 68–74, 2001.
[33] T. W. Bickmore and J. Cassell, "Relational agents: A model and implementation of building user trust," in *Proc. SIGCHI Conf. Human Factors Comput. Syst.*, 2001, pp. 396–403.
[34] J. Cassell and T. W. Bickmore, "External manifestations of trustworthiness in the interface," *Commun. ACM*, vol. 43, no. 12, pp. 50–56, 2000.
[35] M. Feidakis, T. Daradoumis, and S. Caballé, "Emotion measurement in intelligent tutoring systems: What, when and how to measure," in *Proc. 3rd Int. Conf. Intell. Netw. Collaborative Syst.*, 2011, pp. 807–812.
[36] M. Wong, "Emotion assessment in evaluation of affective interfaces," University of Waterloo, Ontario, Canada, 2006.
[37] R. Pekrun, T. Goetz, A. C. Frenzel, P. Barchfeld, and R. P. Perry, "Measuring emotions in students' learning and performance: The achievement emotions questionnaire (AEQ)," *Contemp. Educ. Psychol.*, vol. 36, no. 1, pp. 36–48, 2011.
[38] P. G. Zimmermann, S. Guttormsen, B. Danuser, and P. Gomez, "Affective computing - a rationale for measuring mood with mouse and keyboard," *Int. J. Occup. Saf. Ergon.*, vol. 9, no. 4, pp. 539–551, 2003.
[39] P. G. Zimmermann, "Beyond usability-measuring aspects of user experience," Swiss Federal Institute of Technology Zurich, Zurich, Switzerland, 2008.
[40] A. Burns, et al., "SHIMMERTM - A wireless sensor platform for noninvasive biomedical research," *IEEE Sens. J.*, vol. 10, no. 9, pp. 1527–1534, Sep. 2010.
[41] P. Lewinski, T. M. den Uyl, and C. Butler, "Automated facial coding: Validation of basic emotions and FACS AUs in FaceReader," *J. Neurosci. Psychol. Econ.*, vol. 7, no. 4, pp. 227–236, 2014.
[42] J. C. Lester, S. A. Converse, B. A. Stone, S. E. Kahler, and S. T. Barlow, "Animated pedagogical agents and problem-solving effectiveness: A large-scale empirical evaluation," in *Proc. 8th Word Conf. Artif. Intell. Educ.*, 1997, pp. 23–30.
[43] R. Moreno, R. E. Mayer, H. A. Spires, and J. C. Lester, "The case for social agency in computer-based teaching: Do students learn more deeply when they interact with animated pedagogical agents," *Cogn. Instr.*, vol. 19, no. 2, pp. 177–213, 2001.
[44] M. Courgeon, J.-C. Martin, and C. Jacquemin, "MARC: A multimodal affective and reactive character," in *Proc. Workshop Affect. Interact. Nat. Environ.*, 2008.
[45] R. Moreno and T. Flowerday, "Students' choice of animated pedagogical agents in science learning: A test of the similarity-attraction hypothesis on gender and ethnicity," *Contemp. Educ. Psychol.*, vol. 31, no. 2, pp. 186–207, 2006.
[46] M. Haake and A. Gulz, "Visual stereotypes and virtual pedagogical agents," *Educ. Technol. Soc.*, vol. 11, no. 4, pp. 1–15, 2008.
[47] N. Fourati, A. Richard, S. Caillou, N. Sabouret, J.-C. Martin, E. Chanoni, and C. Clavel, "Facial expressions of appraisals displayed by a virtual storyteller for children," in *Proc. Int. Conf. Intell. Virtual Agents*, 2016, pp. 234–244.
[48] A. Bandura and W. H. Freeman, *Self-Efficacy: The Exercise of Control*. New York, 1997.
[49] P. Steehouder and M. Tijssen, "Opbrengsten in Beeld. Rapportage Aanvalsplan Laaggeletterdheid 2006-2010," CINOP, Den Bosch, 2011.

Dylan G.M. Schouten received the MSc degree in human-technology interaction, focused particularly on user-system interaction experience. He is currently finalizing the PhD thesis on the design and evaluation of VESSEL, a Virtual Environment to Support the Societal participation of Low-literates, as well as researching the use of serious games and novel combinations of top-down and bottom-up assessment methods to unobtrusively assess leadership skill.

Fleur Venneker received the MSc degree in human-centered multimedia. The work done in this study was the basis for her graduation. She currently works as a consultant and user experience designer at the Amsterdam-based company Ebicus.

Tibor Bosse is an associate professor in the Behavioural Informatics Group, Vrije Universiteit Amsterdam. His main research interest is to enhance the believability and effectiveness of Intelligent Virtual Agents by endowing them with dynamic computational models of human behaviour, which are rooted in psychological and social theories. His recent work has an emphasis on the use of IVAs for training of social skills such as aggression de-escalation and cultural awareness.

Mark A. Neerincx is full professor of human-centered computing with TU Delft and principal scientist Perceptual and Cognitive Systems at TNO. His recent research focuses on the situated cognitive engineering of electronic partners (ePartners) that support the social, cognitive, and affective processes in human-automation collaboration to enhance performance, resilience, health, and/or wellbeing. Examples are the Horizon2020 "Personal Assistant for healthy Lifestyle" (PAL) project that develops a physical and virtual robot for children with diabetes, the "IUALL" project on inclusive design and ePartners that enhance citizens' participation in the (local) society, and the "SWELL" project on sensing, modelling and support techniques for stress self-management).

Anita H.M. Cremers received the master's degree in computational linguistics from Tilburg University, The Netherlands, and the PhD degree in natural language human-computer dialogue from the Eindhoven University of Technology, The Netherlands. She is senior scientist with TNO as well as professor ('lector') with the Utrecht University of Applied Sciences (The Netherlands). Her current main fields of expertise are human-computer interaction design and co-design. She focuses mainly on users with limited cognitive and ICT skills.